Assessment of VoIP Quality over Internet Backbones

Athina P. Markopoulou, Fouad A. Tobagi, Mansour J. Karam

Abstract— **As the Internet evolves into a ubiquitous communication infrastructure and provides various services including telephony, it will be expected to stand up to the toll quality standards set by traditional telephone companies. Our objective in this paper is to assess to what extent today's Internet is meeting this expectation. Our assessment is based on delay and loss measurements taken over wide-area backbone networks, considers realistic VoIP scenarios and uses quality measures appropriate for voice. Our findings indicate that although voice services can be adequately provided by some ISPs, a significant number of paths lead to poor performance even for excellent VoIP end-systems. This makes a strong case for special handling of voice traffic on those paths. Even on the good paths, rare loss events can occasionally cause perceptible degradation of voice quality. Finally, the appropriate choice of the playout buffer scheme for each path was found to be of critical importance for the perceived quality.**

I. Introduction

The Internet is evolving into a universal communication network and it is contemplated that it will carry all types of traffic, including voice, video and data. Among them, telephony is an application of great importance, particularly because of the significant revenue it can generate. In order for the Internet to constitute an attractive alternative to the traditional Public Switched Telephone Network (PSTN), it must provide high quality "Voice over IP" (VoIP) services. Our main objective is to assess to what extent today's Internet stands up to these toll-quality expectations. In the process, we identify those aspects that may lead to poor voice quality.

Our approach in addressing this problem has three main characteristics. First, we use delay and loss measurements collected by sending probes between measurement facilities at five different US cities, connected to the backbone networks of seven different providers. These measurements correspond to a large number of paths (43 in total) and a long period of time (2.5 days) and they are rich enough to capture the behavior of Internet backbones. Second, we use appropriate voice quality measures that take into account various transmission impairments. For this purpose, we compile into a single model the results of several studies conducted for specific impairments and we develop a methodology for rating calls. Finally, we take into account the effect of the different components of the VoIP system, with emphasis placed on the playback buffer component.

Although this study is limited to an assessment of Internet backbones, the results obtained are very useful. Indeed, backbone networks are an important part of the end-to-end path (i) for long distance VoIP calls and (ii) for calls that are serviced by a combination of a switched telephone network in the local area and Internet backbones for the long haul. Although backbone networks are usually overprovisioned and cause negligible degradation to data traffic, our study shows that this is not always the case for voice traffic.

Indeed, a large number of the paths performed poorly for VoIP traffic, mainly due to high delay and large delay variability that hurt voice much more than data traffic. Furthermore, if more stringent communication requirements, such as interactivity levels suited for business conversations, are imposed, these paths become totally unacceptable for telephony use. Paths with low delay and low delay variability exhibit in general excellent performance and are appropriate for telephony use. However, even those networks experience occasionally long periods of loss that can affect voice conversations.

As far as the VoIP system is concerned, we consider both fixed and adaptive playback buffer schemes. In both cases, we identify a tradeoff in quality degradation between data loss and increased delay in the buffer, leading to an appropriate choice for the playback delay that takes into account this tradeoff. With regards to adaptive playback schemes, we find that they can adapt to slowly varying delays but not to all the delay spikes that have been observed in the measurements. Furthermore, the problem of tuning the parameters of the adaptive schemes to the delay characteristics experienced on different paths is not an easy one to solve.

The paper is organized as follows. Section II describes the components of the VoIP system under evaluation. Section III presents the quality measures used for assessing the impairments over the network and our methodology for rating a call. In Section IV we describe the probe measurements and classify the traces into categories according to their delay and loss characteristics. In Section V we apply our methodology to the traces, we obtain and discuss numerical results pertaining to phonecalls quality. Section VI concludes the paper.

II. VoIP system

In this section we consider the VoIP system, shown in Fig. 1, we identify and discuss its components.

The first component is the *encoder* which periodically samples the original voice signal and assigns a (usually fixed) number of bits to each sample, creating a constant bit rate stream. The traditional sample-based encoder G.711 uses Pulse Code Modulation (PCM) to generate 8 bits samples per 0.125 ms, leading to a data rate of 64 Kbps. In the same family of sample-based encoders, G.726 uses ADPCM to achieve 16-40 Kbps. Recent frame-based encoders provide drastic rate reduction (i.e. 8 Kbps for G.729, 5.3 and 6.4Kbps for G.723.1) at the expense of additional complexity and encoding delay as well as lower quality.

Further reduction in the data rate can be achieved if no signal is encoded during silence periods, a technique known as Voice

This work was supported in part by RouteScience Technologies Inc. RouteScience also provided the measurements used in this study.

Athina Markopoulou and Fouad Tobagi are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305. Emails: {amarko, tobagi}@stanford.edu. Mansour Karam is with RouteScience Technologies Inc., 159 Second Ave., San Mateo, CA 94401. Email: mans@routescience.com

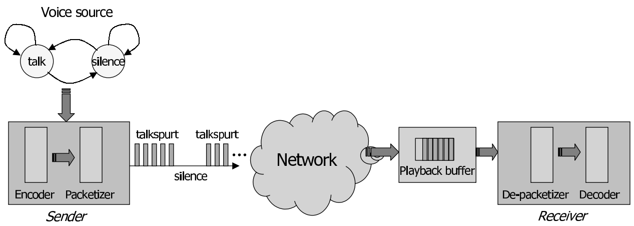


Fig. 1. VoIP System

Activity Detection (*VAD*). It is known that speech can be modeled as a process that alternates between talkspurts and silences that follow exponential distributions with a mean of 1.2 and 1.8 sec respectively, [1]. However, VAD systems tend to elongate the talkspurts by a period called the hangover time, [11]. For the purpose of our simulations, we consider exponential durations with a mean of 1.5sec for both talkspurts and silences, similarly to [16], [17].

The *packetizer* follows the encoder and encapsulates a certain number of speech samples (for G.711) or a certain number of frames (for G.729, G.723) into packets of equal sizes and adds the RTP header (12B). We also take into account the UDP (8), IP (20B) and Data Link headers.

As the voice packets are sent over an *IP network*, they are subject to variable delays and network drops.

An important component at the receiving end, is the *playback buffer* whose purpose is to absorb variations in delay and provide a smooth playout. This is achieved by holding arriving packets until a later playout time in order to ensure that there are enough packets buffered to be played out continuously. Any packet arriving after its scheduled playout time is discarded. Clearly, there exists a trade-off between delay and loss. The playback buffer may operate in one of two modes: fixed or adaptive.

A fixed scheme schedules the playout of a packet after a fixed (network and buffering) delay from its sending time, the same for all packets. The value of this fixed delay is important in order to avoid either unnecessarily delaying or dropping of packets. It should be chosen based on some knowledge of the delay on the path. However, such an assessment may not always be possible or the statistics of the network delay itself may change with time. In addition, a fixed playback scheme needs synchronization between the source and the receiver in order to guarantee the chosen end-to-end fixed delay.

For these reasons, extensive work, [25], [23],[21], is being conducted on adaptive playout schemes that dynamically adapt the playout time to closely follow the variations in network delays. How often one might need to adapt depends on how fast the delay characteristics change on the path. A simple yet effective scheme has been studied in [25]; it decreases both delay and loss by adapting at a short time scale, namely at the beginning of each talkspurt. A more sophisticated scheme that adjusts the playout rate in the middle of a talkspurt without the user perceiving it, is described in [21].

In our study we considered both fixed and adaptive schemes. A fixed scheme with an appropriate choice of delay is useful as a benchmark for the assessment of a path. We also implemented the adaptive schemes proposed in [25], and we used the “spike-detection” as our baseline scheme. This algorithm learns from the delay experienced by previous packets, updates the moving averages of the mean d_{av} and the standard deviation v of network delay, and adapts the playout time at the beginning of each talkspurt to be $p = d_{av} + 4v$. It also performs delay spike detection and adapts faster to the network delays within a spike. We used the default parameters of [25] and 30 ms as the delay of the first packet, the nominal value used in [2]. We did not allow decrease in the playout time of a talkspurt that would overwrite already buffered talkspurts. The objective of this paper is not to design a new playback scheme or to exhaustively evaluate all existing ones, but it is instead to use realistic schemes to evaluate VoIP performance.

The playout buffer delivers a continuous stream of packets to the *depacketizer* and eventually to the *decode*r which reconstructs the speech signal. Decoders often implement Packet Loss Concealment *(PLC)* that produces a replacement for a lost packet, similar to the original one, by filling in silence or noise, by interpolating or even by regenerating the packet from the surrounding ones. Error concealment works best for small loss rates and durations. The reader is referred to [24] for details on packet loss recovery techniques for streaming audio in general.[1]

Each of the above components along the path of the packetized voice, may introduce delay and loss. The components of the end-to-end delay are the following (i) encoding and packetization delay at the sender (ii) propagation, transmission and queuing delay in the network and (iii) buffering and decoding delay at the receiver. Distortion of the original voice signal may occur: (i) at the low rate encoder (ii) in the network due to loss and finally (iii) at the receiver due to drops in the playback buffer.

Another important impairment, omitted for simplicity from Fig. 1, is *echo*, the reflection of the participants’ signals, perceived as delayed and attenuated versions of their own voices. The larger the end-to-end delay, the more annoying is the echo. Although one might at first think that echo cannot happen in a packetized voice system, reflections may indeed happen (i) at the four-to-two wires hybrid connection between a packet and a circuit switched network and (ii) at the PC end-point when the microphone picks up the remote person’s voice from the speaker as well as multiple reflections in the room and bounces them back. Both types of echo can be controlled by an Echo Canceller, that should be located as close to the source of echo as possible. The reader is referred to [20] and [29] for more details.

III. VoIP Quality Assessment Methodology

With our end-goal being the assessment of VoIP performance over today’s Internet, we first need to choose quality measures relevant to voice traffic. There are several sources of impairments, identified in Section II. Network performance is usually presented in terms of delay and loss statistics. However, the ultimate judge for the quality of a phone conversation is the user and the most appropriate quality measure is the user’s opinion. A commonly used subjective metric is the *Mean Opinion Score* (MOS), i.e. the average of ratings on a scale from 1 to 5, given

[1] Although not evaluated in our study, it is worth mentioning that actual audio tools, such as [28], may include additional error resiliency mechanisms. These may include transmission of layered or redundant (FEC) audio, interleaving frames in packetization, retransmissions, communication between sender and receiver in order to switch encoders or data rates.

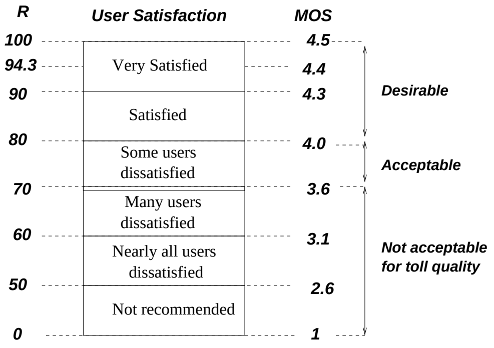


Fig. 2. Voice quality classes

by individuals under standardized conditions.

Numerous studies over the last decades have performed subjective tests to quantify the effect of individual impairments on conversation quality. They map some measurable expression of loss ([11], [3], [7], [30]) or delay([19] and [15]) to a single MOS rating, by means of statistical analysis of subjective tests results. In Subsection III-A, we combine the data provided by the above studies using the Emodel computational model, [12][13][14], to get a single MOS rating for a speech segment. In the process, we confirm the consistency among the results of these different studies and thus their validity. In Subsection III-B, we combine recent studies, [9], [10], [4], [5],[6], to develop a methodology to rate an entire voice call, consisting of multiple short speech segments.

A. Voice Quality Measures

The Emodel is a computational model, standardized by ITU-T in [12][13][14], that uses transmission parameters to predict the subjective quality of packetized voice. We use it to combine individual delay and loss impairments into a single rating R on a scale from 0 to 100, which can be further translated into MOS. User satisfaction, and the corresponding R and MOS ranges, are shown in Fig. 2. The operational range for PSTN voice quality corresponds to $MOS \geq 3.6$. The desirable range of operation for toll quality is $MOS \geq 4$.

The Emodel combines different impairments based on the principle that the perceived effect of impairments is additive, when converted to the appropriate psycho-acoustic scale (R).

$$R = (R_o - I_s) - I_d - I_e + A \quad \textit{\textbf{(1)}}$$

The details of equation (1) are as follows. Both Ro (effect of noise) and Is (accounting for loud connection and quantization) terms are intrinsic to the voice signal itself and do not depend on the transmission over the network. Thus, they are irrelevant for the purpose of comparing VoIP to PSTN calls. I_d and I_e capture the effect of delay and signal distortion respectively and they are discussed below, in a separate subsection each. A stands for the advantage factor that captures the fact that users might be willing to accept some degradation in quality in return for the ease of access, e.g. using cellular or satellite phone. For the purpose of comparison to PSTN calls, this factor is set to 0.

A.1 Delay impairment I_d.

The I_d factor models the quality degradation due to one-way or “mouth-to-ear”(m2e) delay. Id can be further broken into three terms:

$$I_d = I_{dte}(m2e, EL_2) + I_{dle}(m2e, EL_1) + I_{dd}(m2e) \quad \textit{\textbf{(2)}}$$

The terms $I_{dte}(m2e, EL_2)$ and $I_{dle}(m2e, EL_1)$ capture the impairments due to talker and listener echo respectively. EL_1, EL_2 are the echo losses in dB at the points of reflection and their value depends on the echo cancellation used. $EL = \infty$ (infinite echo loss) corresponds to perfect echo cancellation. $EL = 51\,dB$ corresponds to a simple yet efficient echo controller. The third term $I_{dd}(m2e)$ captures the interactivity impairment when the m2e delay is large, even with perfect echo cancellation. Indeed, large m2e delay may lead to “collisions” when participants talk in the same time, or may force them to take turns and thus take longer to complete the conversation. (2) is also in accordance with ITU recommendation G.114, [15], which provides specifications for one-way transmission time. According to (2), $m2e$ delays below 150 ms should not affect interactivity, a claim that motivated us to further investigate this point. There is indeed a dimension that is not captured by (2), that of the different modes of conversation or “tasks”.

“Tasks” are defined in [19] to be types of conversation with different switching speed and thus different sensitivity to delay. For example a business call might involve shorter messages and higher speed in switching among participants, than a social call. The fact that the Emodel does not account for tasks, implies that the I_d curves provided hold for the average of all tasks used in subjective tests. [19] assumes $EL = \infty$ and studies the effect of delay on six types of tasks. The most stringent one is “Task 1”, where people take turns reading random numbers as quickly as possible. On the other extreme, “Task 6” is the most relaxed type, free conversation. Business calls are more likely to have the stringent requirements of the first tasks.

We take into account the data provided by [19] in evaluating the loss of interactivity. We use the echo impairment terms as provided by the Emodel. The combined curves, [2] that capture the total delay impairment, are shown in Fig. 3.

A.2 Loss impairment Ie

The Ie term in equation (2), called the “Special Equipment Impairment factor” in the context of Emodel, captures the distortion of the original voice signal due to low-rate codec, and packet loss in both the network and the playback buffer. Table I gives the intrinsic quality, and thus the I_e, in the absence of packet loss, for various encoders. G.711 starts at the highest intrinsic quality (94.3). Modern encoding schemes, such as those used by G.729 and G.723.1, achieve higher compression at the expense of lower intrinsic quality, which makes them less tolerable to loss during their transmission. The distortion as a function of packet loss also depends on whether or not PLC is implemented. Roughly speaking, the impairment increases by about

[2] The delay impairment curves show $I_d > 0$ even in the $[0, 150ms]$ range where it is usually shown to be 0, [12],[15],[29]. This is because we preferred to linearly interpolate among the data points provided by [19] and be on the conservative side, than to assume $Id = 0$ over that range. This interpolation might be above the real I_d at most by 10 points in the R-scale, which is small anyway.

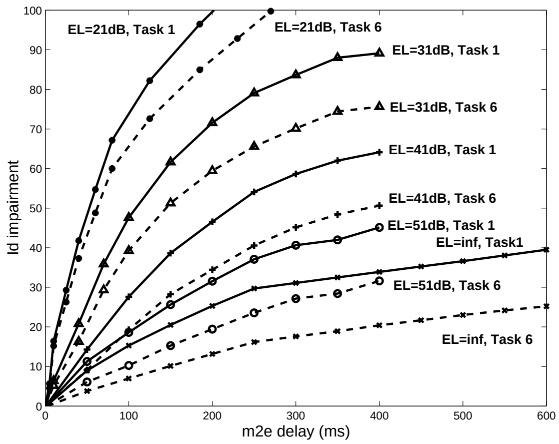


Fig. 3. Delay impairment (I_d) as a function of the mouth-to-ear (m2e) delay. Parameters: (i) the type of Task $\{1, 2, ..., 6\}$ and (ii) the Echo Loss (EL) in dB.

TABLE I
STANDARD ENCODERS, WITH KNOWN I_e

Standard	Codec	Rate	I_e	R_{intr}
	type	(Kbps)	(loss=0)	
G.711	PCM	64	0	94.3
G.729	CS-ACELP	8	10	84.3
G.723.1	ACELP	5.3	19	75.3
G.723.1	MP-MLQ	6.3	15	79.3

4 units in the R scale per 1% packet loss for codecs with PLC and by 25 units for codecs without PLC. However, some type of packet loss concealment (PLC) is a common practice today; it is built-in in G.723.1, and G.729 and it can be added for G.711.

Fig. 4 shows how the I_e impairment increases with the packet loss rate for different codecs, packet sizes and PLC techniques. The curves provided by the Emodel are shown in solid lines. The following packetization is considered: a G.711 packet contains $10\,ms$ of speech; a G.729-A packet contains two frames ($10\,ms$ each); a G.723.1-A packet contains one frame ($30\,ms$). All the curves, but one, assume uniform packet loss. The curve for bursty loss is based on the AT&T contribution [3], which used a two-state bursty loss model and a maximum loss duration of 100 ms. We are particularly interested in the bursty loss which is the case in the Internet traces.

In addition to the above curves, we consider results from other studies for the purpose of checking the validity of the Emodel I_e curves as well as increasing our evaluation options. Results concerning G.711 can be found in [3], [11], [7]; results on G.723.1 can be found in [30].

Cox and Perkins in [3], studied the effect of both uniform and bursty frame erasure on G.711 with Frame Erasure Concealment and a frame of 10 ms. Later, their study evolved into the Emodel curves for G.711. Also, the ETSI Tiphon project, [8], collected contributions of subjective results on the effect of IP packet loss, delay and echo, which are not all included in [14]. Early on [11], Gruber applied uniform loss (of various rates and durations) on PCM speech and obtained MOS ratings. Those results are comparable to the Emodel curve for G.711 without PLC. We convert the MOS ratings from [11] for loss durations of 10 ms into I_e impairments, we plot them in dotted line on Fig. 4 and we confirm that they agree with the I_e provided by the Emodel. Results from [11] for longer loss durations ($16\,ms$, $64\,ms$, $256\,ms$) and rates from 0.5 to 20% show huge loss impairments because they are obtained without PLC, which is an unrealistic choice in the context of today's VoIP.

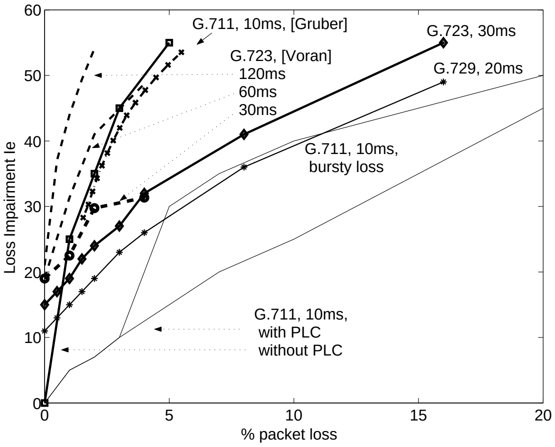


Fig. 4. Loss impairment (I_e) as a function of the packet loss rate. Parameters to be specified for each curve: (i) Standard (G.711, G.729-A, G.723.1-A) (ii) speech duration in one packet (iii) use of PLC (iv) source of data (Emodel [13], Gruber [10], Voran [26], Cox & Perkins [21]).

In [30], Voran applied various impairments (various rates, durations and types of temporal discontinuities) on G.723.1 encoded speech, with VAD and PLC, a frame of 30ms and a rate of 5.3Kbps. We translate the degradation in MOS into a loss impairment value I_e. First, we plot this Ie for G.723.1 and 30ms in Fig. 4 and we observe that it quantitatively agrees with the Emodel curve, for the loss rates range of $[0, 4\%]$. The small deviation is due to the different encoding schemes considered for G.723.1 by the Emodel (MP-MLQ) and by [30](ACELP), thus the small difference in intrinsic quality. Second, we translate the MOS values provided by [30] for $30\,ms$, $60\,ms$ and $120\,ms$ gap durations, into the equivalent I_e ratings, and we plot them in dashed line in Fig. 4.

B. VoIP call quality

The previous subsection provides a rating for a segment of packetized speech that incurred a certain packet loss and delay. This is appropriate for rating short speech samples, like those used in the subjective tests that led to the above curves , i.e. in the order of a few seconds for I_e and in the order of 1 min for I_d. However, this approach is not applicable to entire phone calls, lasting several minutes. Calculating the average loss rate and the average delay over the entire phonecall would only give a rough estimate.

A natural approach is to divide the call duration into fixed time intervals and assess the quality of each interval independently, using the I_d and I_e curves of Subsection III-A. Independent $MOS(t)$ rating of each short interval t has been shown in [9] to correlate well with the continuous instantaneous rating of the call. Evaluating each interval in terms of I_e leads to transitions between plateaus of quality, as shown in dashed line in Fig. 5.

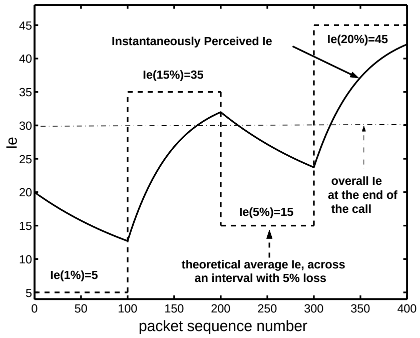


Fig. 5. Transitions between periods of high and low loss. Theoretical vs. instantaneously perceived I_e.

However, transitions between high and low loss periods are perceived with some delay by the listener, as opposed to abrupt changes between plateaus. For example in Fig. 5, a human would perceive and rate the changes in quality using the smooth solid line instead of the dashed one. Therefore, a model monitoring quality over time should take into account time constants. [10] demonstrated this "recency effect" and noticed that it takes longer for a subject to forget transitions to bad than to good quality. Instantaneously perceived I_e is considered by [4] to converge toward the $I_e(loss)$ for a gap or burst, following an exponential curve with time constants $T_{bad} = 5\,sec$ for the high loss and $T_{good} = 15\,sec$ for the low loss periods.

In addition, there is no guarantee that the assumption of uniform loss, underlying the I_e curves, holds for Internet traces. To appropriately handle the burstiness in packet loss, [4] and [7] proposed the use of variable length intervals to calculate I_e over them. More specifically, they defined high and low loss periods, called "bursts" and "gaps" respectively.[3] The use of variable intervals appropriately addresses the burstiness in the following ways. First, the loss during gaps is enforced to be uniform by the definition of a gap. As for the burst periods, we decided to use the curve of Fig. 4 for bursty loss. Second, by dynamically partitioning each trace into its own gaps and bursts, we emphasize the periods of high loss, as opposed to calculating the loss rates over arbitrarily long intervals and smoothing them out.

It has also been shown, [9], that the rating an individual would give at the end of a call is captured at a first approximation by the time average of the instantaneously perceived MOS. [4] further adjusted the final rating to include the effect of the last significant burst and demonstrated good correlation with subjective results, [5], [6]. Notice however, that an individual might forget some bad moments in the middle of the call, that a network provider might be interested in monitoring and eliminating. Therefore, in our assessment of an entire call, we use not only the rating described in [4] to simulate the opinion of an individual, but also the worst quality experienced during a call, in order to highlight bad events.

In summary, our approach for rating an entire call is the following. We use the idea of bursts and gaps from [4] and [7] to address the burstiness. However (i) we avoid the computational simplifications used in [4] to decrease the processing time and provide an online service and (ii) we use the bursty loss curve for the high loss periods. We also use the concept of perceived quality from [10]. As for the rating of an entire call, we consider both the -lenient- rating of [4] at the end of the call and the worst instantaneous MOS during the call. We also differ from the previous approaches in that we consider talkspurts and silences. Finally, we studied (but omitted from this paper for lack of space) the sensitivity of our approach to parameters such as the gap length (g_{min}), the time constants of the "recency effect" and different functions for calculating an overall MOS from an instantaneous $MOS(t)$.

It is worth mentioning that some commercial systems for online monitoring of VoIP quality are currently being developed along the same lines. The authors are aware of two such tools: (i) one by Telchemy, [4] and (ii) another by NetIQ, [31].

IV. Internet Measurements

In this section we describe the measurement experiment and the delay and loss characteristics of the traces collected.

A. Description of probe measurements

Our study is based on delay and loss measurements provided by RouteScience Inc. Probes were sent by and collected at measurement facilities in 5 major US cities: San Jose in California (SJC), Ashburn in Virginia (ASH), Newark in New Jersey (EWR), Thornton in Colorado (THR) and Andover in Massachusetts (AND). 43 paths in total were used, obtained from seven different providers, which we refer to as $\{P_i\}_{i=1}^{i=7}$ for anonymity purposes. The measurement setup is shown in Fig. 6. E.g. the bidirectional arrow drawn between SJC and AND means that measurements were collected from SJC to AND and from AND to SJC using providers P_3 and P_6. All paths are backbone paths, connected to the measurement facilities through either T3 or T1 links. Paths for all providers are two ways, except for those shown in parenthesis.

The probes were 50 Bytes each and were sent every $10\,ms$ from Tuesday 2001/06/27 19:22:00 until Friday 2001/06/29 00:50:00 UTC. GPS was used to synchronize senders and receivers and the network delays were inferred by subtracting the sender from the receiver timestamp. The load generated by the probes was insignificant and did not affect the delay and loss characteristics of the networks.

By taking into account the providers' access bandwidths we are able to compute the transmission time and infer delays for any voice packet size from the probe delays.[4] The 10ms sending interval is small enough to simulate the highest rate a VoIP encoder/packetizer might generate packets at. By appropriately

[3] More specifically, if the number of consecutive received packets between two successive losses is less than a minimum value g_{min}, then the sequence of the two lost packets and the intervening received packets is regarded as part of a burst ; otherwise, part of a gap. We use $g_{min} = 64$ packets which results in gaps and bursts of meaningful durations in the order of 0.5-1 sec, and matches well the loss patterns in our traces.

[4] For example, a G.729 packet containing one frame generated every 10 ms (8 Kbps rate) has exactly the size of a probe: 10B for the payload and 40B for the IP/UDP/RTP header. A G.711 packet sent every 10ms at 64 Kbps, contains 80B (payload) + 40B (header) = 120B, which is longer than the probe by 70B. The transmission of 70B takes $0.012ms$ and $0.038ms$ over a T3 and a T1 access link, respectively. We did subtract these differences in delays, which are anyway negligible compared to the network delays. The difference in transmission times inside the backbone (bandwidth in the order of 100 Mbps-1 Gbps) are even shorter and thus ignored.

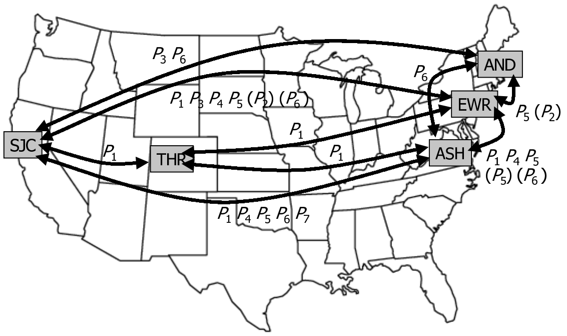
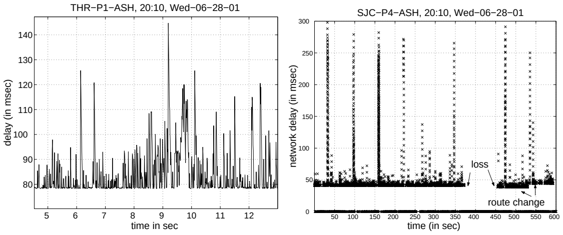


Fig. 6. Probes measurements



(a) Example trace, type E

(b) Example trace, type D

Fig. 7. Example patterns for two different providers

omitting some probes we can simulate lower packet rates or silence periods.[5]

B. Traces description

We classify the 43 paths into five types, based on their fixed (i.e., propagation and transmission) and variable (i.e. queuing) components of the delay. Table II shows an example of each type. Paths of type A and B connect ASH, EWR and AND on the east coast and have low propagation delays, i.e. below 10ms. Paths of type C, D on E on the other hand, connect cities across the US. We further distinguish, based on the variable component of delay. Paths of type A and C have practically no queuing (as indicated by the delay percentiles which are close to the fixed component); they turn out to be the best for carrying VoIP. Paths of type B and D have in general low queuing, except for clustered delay spikes (which last 1-2 sec each and appear every almost 70sec, see Fig. 7(b)), that lead to delay percentiles, 4-5 times higher than the fixed delay, in Table II. Finally, paths of type E are coast-to-coast loaded paths. The queuing component is high and the delay varies slowly in a short time scale (see Fig. 7(a) and high delay percentiles in Table II), as well as across the day (see the significant increase during business hours in Fig. 8).

We observe that network loss events of various durations are spread across all types of paths.

- Only 3 out of the 43 paths had consistently *no loss* during the 2.5 days observed. The rest of them incurred loss durations that varied from 10ms up to 33.72sec, although the *average* loss *rates* were *low* (e.g. <0.2%).

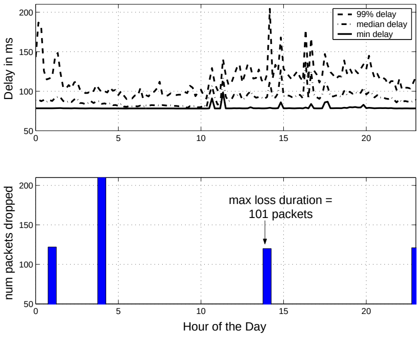


Fig. 8. Example path of type E (THR-P_1-ASH) across an entire day (Wednesday 06/27/01).

- 6 out of 7 providers experienced *outage periods* 10-220sec for 1-2 times per day. For two of these providers, these outages were correlated with changes in the minimum delay even for a few ms, as in Fig. 7(b). We attribute these events to *routing changes:* the propagation delay changes and there is loss for the time required by routing protocols to converge. For one provider, this event was a recurrent phenomenon (3-4 times per day). For the reasons behind the rest of the outages, we speculate link failures or maintenance at night time.
- 0.5-2sec loss durations were correlated with delay spikes.
- The number of out-of-order packets was negligible.

An important observation is that paths of the same provider have the same consistent delay variability and loss pattern, whether they are short or long distance. This is intuitively expected as a backbone is shared by many paths of the provider. For example, each provider experienced long loss durations (5-33sec) on many paths simultaneously, hinting to a failure on a backbone link. All P_3 paths experience single (10ms) losses at 0.2% rate. All P_4 paths periodically exhibit clusters of high spikes, Fig. 7(b), and belong to the categories B or D . All paths of type E belong to provider P_1.

V. Numerical Results

In this section we apply the methodology of Section III to the traces of Section IV. In doing so, we first go through the analysis of an example path. Then, we present results for all types of paths.

A. Example path

Let us first consider the example trace of type E and a call taking place from 14:00 until 14:15 on 06/27/01. The selected trace exhibits large delay variations and a period of sustained loss. Fig. 9(a) shows the network delays and the playout times and Fig. 9(b) shows the corresponding perceived quality. Let us first consider a fixed playout, e.g. 100ms. Clearly, the larger the playout delay, the larger the delay impairment I_d but the smaller the loss and the loss impairment I_e. The overall MOS is a combination of both I_e and I_d according to equation (1).

Clearly, there exists a tradeoff between loss and delay, shown in Fig. 10, and a value of the playout delay that maximizes MOS. The optimal fixed $m2e$ delay for the example call is around $200\,ms$ and results in $MOS = 4$. The [25] adaptive scheme operated near the optimal region achieving

[5] For example, by omitting 100 consecutive probes, we simulate a silence period of $100 \cdot 10ms = 1sec$. Also, by omitting every other probe packet, we can simulate voice packets sent every 20ms.

TABLE II

TYPICAL PATHS

Type	**Num**	**Example path**				**Delay (in msec)**				**Loss**			
										usual loss events		**long outages**	
		From	Prov.	To	Dist.	min	50%	98%	99%	avg clip (msec)	#clips per hour	duration (sec)	times per day
A	11	EWR	P_6	ASH	short	3.4	3.6	3.7	3.72	20	1-5	5-12	1-2
B	2	ASH	P_4	EWR	short	6.8	7.2	120	200	20	2-3	12-25	1
C	16	SJC	P_5	EWR	long	32.7	32.8	33.5	34.5	0	0	2	1
D	4	SJC	P_4	ASH	long	45.1	45.4	170	225	10	1-2	15-25	2-3
E	10	THR	P_1	ASH	long	77.8	78.2	100	210	10	2-20	1	1

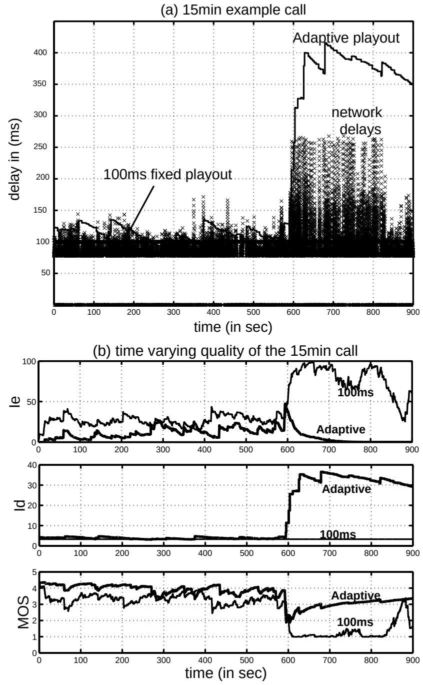


Fig. 9. An example of 15min call (14:00-14:15, Wed. 06/27/01, THR-P_1-ASH). Both fixed and adaptive playout considered. (a) Network and playout delays (b) Resulting I_d and I_e impairments and instantaneously perceived MOS.

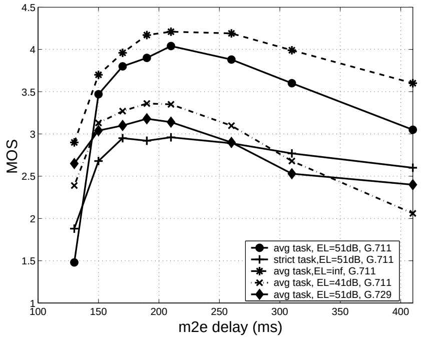


Fig. 10. Delay-loss tradeoff for the example call, considering various VoIP parameters (average or strict (Task 1) task, $EL = \{\infty, 51dB, 41dB\}$, G.729 or G.711). "MOS" refers to the overall rating at the end of the 15min call. Fixed playout is applied during the entire call.

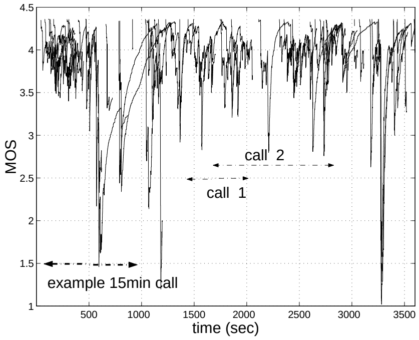


Fig. 11. Time varying quality of (50) calls, over an one-hour period, on the path THR-P_1-ASH.

$max\ MOS = 3.6$ for an average delay of $122\,ms$. This performance was achieved using a favorable VoIP configuration, i.e. G.711 encoding (which has a high intrinsic quality), an adequate echo cancellation ($EL = 51dB$) and a medium interactivity requirement.

A similar loss-delay tradeoff holds under any VoIP configuration. However the optimal delay range as well the maximum achievable MOS may differ. For example, G.729, which starts at a lower intrinsic quality, can achieve a max $MOS = 3$ and thus cannot be carried at acceptable quality levels during the 15 minutes considered period. Similarly, a strict interactivity requirement (e.g. "Task 1") or an acute echo (e.g. $EL = 41\,$dB), would lead to $max\ MOS \cong 3$, which is unacceptable.

Having discussed one call in detail, let us now consider many calls initiated at random times, uniformly spread over an entire hour, e.g. from 14:00 to 15:00. We consider exponentially distributed call durations as in [26]. 150 short ($3.5\,min$ mean) and 50 long ($10\,min$ mean) durations simulate business and residential long distance calls, respectively. Fig. 11 shows the instantaneous quality of some of these calls, that varies with time. To rate each call, we use both the minimum MOS during the call (that a network operator might want to eliminate) or the more lenient rating at the end (that a human would give), as discussed in length in Section III-B. Fig. 12 shows the cumulative distribution (CDF) of ratings for the 200 calls, using both measures. If fixed playout is used, Fig. 12(a), then the choice of the fixed value becomes critical: $150\,ms$ is acceptable (only 6% of the calls have final rating below 3.6 and only 8% of them experience a period of $MOS < 3.6$) while $100\,ms$ is totally unacceptable

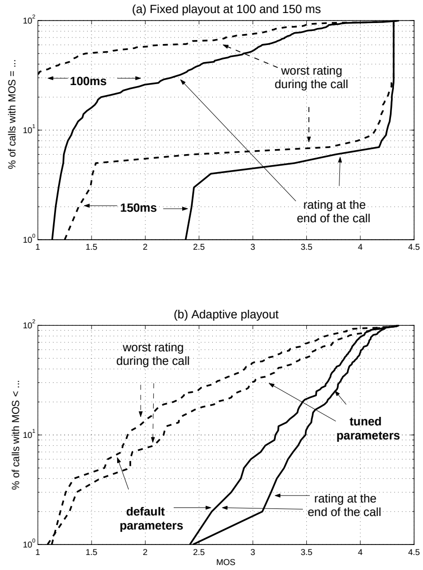


Fig. 12. CDF of call ratings in one-hour period (Wednesday 06/27/01, 14:00-15:00) on a path of type E (THR-P1-ASH).

(90% of the calls have rating at the end below 3.6). For the adaptive playout, Fig. 12(b), we observe the following: (i) the CDF is more "linear" than for the fixed scheme (ii) this performance is acceptable but still not excellent (10% of the calls have overall rating $MOS < 3.6$ and 50% of them experience a period of $MOS < 3.5$ at least once) (iii) tuning of the parameters does not lead to significant improvement.

While in Fig. 12 we plot the entire CDF, in Fig. 13 we consider only some percentiles (i.e. worst rating, 10%, 50%, 90%, 100%) of call ratings for each hour-bin of the entire day. E.g. the points in Fig. 13(a) for $Hour = 14$ are consistent with Fig. 12(a): out of the 200 calls between 14:00 and 15:00, the worst rating was 1.1, 10% of the calls had $MOS \leq 1.4\%$, 50% of the calls had $MOS \leq 3$, 90% of the calls had $MOS \leq 3.75$ and some calls have perfect rating.

Fig. 13(a) shows that a fixed playout at $100\,ms$ is unacceptable when the delays on the path are high, i.e. during the business hours, see Fig.8. A fixed value at $150\,ms$, Fig. 13(b) is a safe choice as no more than 1-2% of the network delays (Fig. 8) exceed it. The bad rating at 14:00 is due to the network and not due to buffer loss. On the other hand, the adaptive playout, Fig. 14(c), had the same performance for the entire day including the business hours, because it was able to adapt to the network delays. However, it did not perform particularly better: 10% of the calls in any hour had $MOS < 3.5$.

B. All paths

We apply the same procedure to the rest of the example paths.

We observe that paths of low delay and low delay variability, of both short (type A) and long (type C) distance, achieve an excellent MOS at all times except for the rare cases when long network drops occur. A high fixed playout delay of 100 ms is sufficient to yield excellent performance. Fig. 14 shows that 90% of the calls on the example path of type A have $4 \leq MOS < 4.4$. The low ratings at 4:00 and 6:00 are due to long network drops of 3 sec and 6 sec respectively. We also observe that the performance degrades when the adaptive playout tries to follow closely the network delay; this is unnecessary in this case that the delay does not vary significantly. Similar findings hold for the paths of type C, which are long distance paths but with delay significantly below $100 - 150\,ms$ and with low variability.

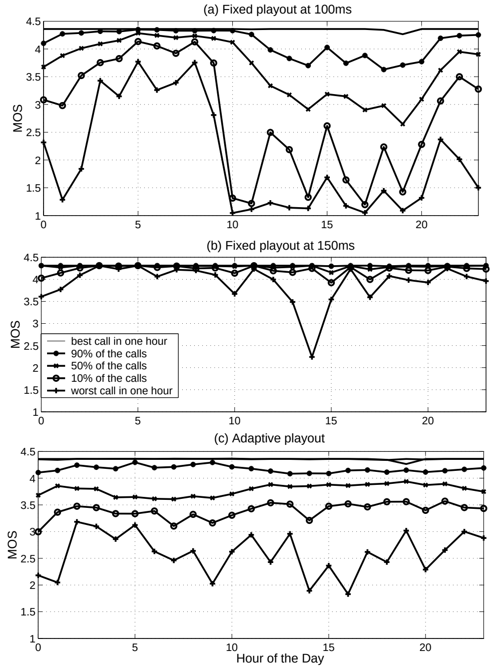


Fig. 13. Call quality statistics for every hour of an entire day (Wednesday 06/27/01) on a path of type E (THR-P_1-ASH). Playout used: (a) Fixed at 100ms (b) Fixed at 150ms (c) Adaptive with default parameters.

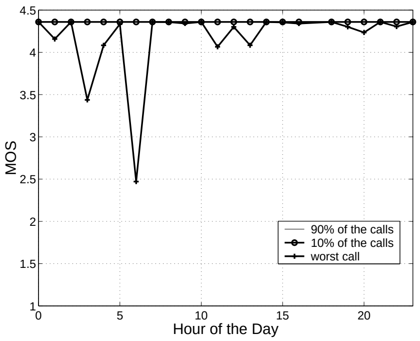


Fig. 14. Call quality on a path of type A (EWR-P_6-ASH) on Wednesday 06/27/01. Fixed playout at 100ms.

In contrast, paths of type B and D exhibit periodically clusters of high spikes, as in Fig. 7(b). Packets from these clusters are dropped at the playout buffer, whether a fixed or the baseline adaptive playout is used. Because delays on these paths do not vary across the day, it makes sense to look at one typical hour. If adaptive playout, Fig. 15, is used with its default parameters, 20% of the calls have overall $MOS < 3.5$, which is unacceptable. Even worse, 80% of the calls experience $MOS < 3.5$ for some period. If a strict interactivity requirement is applied, then the entire CDF degrades by approximately 0.8 unit of MOS. Performance can improve if an appropriately high fixed delay is chosen: only 10% of the calls have overall $MOS < 3.5$.

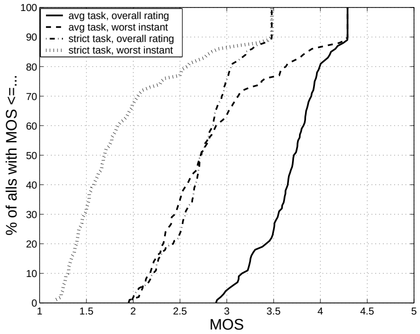


Fig. 15. Call quality for a path of type B (SJC-P4-ASH), on Wednesday 06/27/01 from 20:00 to 21:00. Adaptive playout with default parameters used.

C. Discussion

In this section we discuss the numerical results, we provide some recommendations and directions for future work.

C.1 On the performance of the backbone networks

Our results indicate that some ISP backbones (i.e. those that are over-provisioned and have low delay variability, namely types A and C) are indeed able to provide high quality VoIP today. This is true for both short and long distance paths. In their case, the only problem is the rare occurrence of long periods of loss. This makes a case for identifying VoIP traffic as such, for the purpose of treating it favorably during those rare events (e.g. routing changes).

On the other hand, highly loaded paths (type E) as well as some over-provisioned paths exhibiting frequent delay spikes (types B and D) have poor VoIP performance. Under the best scenarios (namely G.711 encoder, good echo cancellation and low interactivity requirements) these paths are barely able to provide acceptable ($MOS > 3.6$) VoIP service, far below the guarantees of the telephone network. Performance is even worse for stringent application requirements or less favorable system configurations. For example, strict interactivity requirements (Task 1) decrease MOS by roughly 0.5-1 units. Inadequate echo cancellation (e.g.. $EL = 41dB$ instead of $EL = 51dB$) has a similar effect. Support of G.729 , which has lower intrinsic quality, is possible only on paths of types A and C at acceptable quality levels.

The poor VoIP performance on loaded backbones (type E) makes a strong case for separating voice and giving it priority over other traffic in these networks, [18]. The poor performance on paths periodically exhibiting spikes (types B and D) also needs a more sophisticated handling than over-provisioning.

Our observations were similar for both short (A or B) and long distance paths (C, D, E). The reason for this, is that most of these backbone paths have delay not significantly higher than 150ms. Calls going through multiple backbones or through wireless/access networks would incur even larger and more variable delay and even worse performance.

A defining factor for the perceived performance turns out to be the delay variability on a path. Furthermore, we observed that each provider has its own "signature" on a trace, i.e. the same consistent delay and loss patterns. It is very important that the delay pattern on a path be handled by the appropriate playout buffer at the receiver, as discussed in the next section.

C.2 On the Playout Buffer.

Our intention was to consider some realistic playout schemes, as part of the end-to-end VoIP system under evaluation. We first considered fixed playout for a range of fixed playout delays and then a baseline adaptive scheme, [25]. The study of the fixed playout provides a benchmark for comparison.

There exists a tradeoff between delay and buffer loss, Fig. 10, and a maximum $MOS(loss, delay)$ corresponding to the best possible performance on the path. An appropriate choice of the fixed playout buffer is the one that leads to maximum MOS. A good adaptive scheme should also operate around that maximum MOS. As shown in Fig. 10, the maximum MOS is more sensitive to an increase in loss rather than to an increase in delay. The reason for this is that the underlying I_e curves, Fig. 4, are sharper than the I_d curves, Fig. 3. This is why a conservative choice of a (high) fixed playout value prevented packet loss and led to good performance on low delay (i.e. $< 150\,ms$) paths.

The need for adaptive playout comes when (i) the delay is high (close to or above the interactivity constraint of $100 - 150\,ms$) and there is no margin for overestimating it and (ii) when the delay is unknown and the receiver does not know how to select an appropriate fixed value. An adaptive scheme learns, predicts and follows the network delays as closely as possible, thus keeping both delay and loss low. The adaptive playback we considered, [25], was useful on the loaded network (type E) that exhibited high and slowly varying delays but failed on paths of type B/D and A/C. This bad performance can be attributed to (i) the tuning of its parameters and to (ii) the failure to predict the actual delays.

As far as the tuning is concerned, the default parameters, namely the *weights* used for the calculation of the moving averages and the *thresholds* used for spike detection, were optimized for the specific network traces considered in [25]. A single tuning of these parameters that works well for all traces is not an easy (or not even a feasible) problem to solve.[6] Furthermore, even if these parameters are appropriately tuned for a specific network path, the characteristics of the path may change in time. Then the adaptive algorithm may need to adapt its own parameters to match not only the delay pattern of a specific trace but also the change of this pattern in time. We did experiment with these parameters on our traces and achieved roughly reasonable loss rates of 2-4% during an entire call. However, the loss rates during shorter intervals were occasionally much higher.

As far as the delay estimation mechanism of [25] is concerned, it has the following weaknesses. First, the TCP-like prediction ($p = d+4v$) tends to over-estimate delays beyond what is appropriate to preserve interactivity. Second, as also noticed in [23], adapting at the beginning of talkspurts fails to react to short lived spikes while it still unnecessarily leads to high delays.[7] Finally,

[6] For example, delay spikes might be in the order of 100ms for one one trace, and in the order of 10ms for another. A threshold for spike detection at 10ms, would make the 1st trace operate in the SPIKE mode all the time, while choosing it at 100ms would detect no spikes at all in the 2nd trace.

[7] G.729B VAD uses a dynamic hangover scheme, leading to shorter talkspurt and gap lengths on average, which would give the spike detection algorithm more chance to adapt to spikes. So, using 1.5sec average for both G.711 and G.729 might not be fair. G.729B has not been considered in this paper.

trying to closely follow the delays after exiting a spike, often leads to under-estimation and thus loss at the beginning of the next talkspurt, which is particularly difficult to conceal.

This paper did not intend neither to invent new playout algorithms nor to compare all the existing ones. However, in the process of evaluating the end-to-end VoIP system, considering some popular adaptive algorithms and tuning their parameters, it became clear that the appropriate choice of playout scheme for each path is a defining factor for the end-to-end quality. An adaptive algorithm has the potential to perform at least as well as a fixed one, but this is possible only if its mechanisms are carefully tuned to match the network path. This experience further motivated us to continue this work [22] towards designing a playout buffer that would maximize the voice perceived quality $MOS(delay, loss)$ as opposed to delay and loss percentiles. Our scheme: (i) explicitly accounts for the delay impairment by including it in the objective function (ii) adapts slowly to the variations of delay in time but (iii) conservatively over-estimates delay to avoid unnecessary buffer loss, whenever this is allowed by interactivity constraints.

VI. Conclusion

In this paper, we assess the ability of Internet backbones to support voice communication. We consider a realistic configuration of the end-to-end VoIP system. We compare and combine results from various subjective testing studies and we develop a methodology for assessing the quality of a call in terms of relevant measures. Key asset in our study is the use of network measurements collected over backbones of major ISPs.

In general, backbone networks are over-provisioned and thus expected not to be the bottleneck on the path of a flow. Although this might be the case for data traffic, this is not always the case for VoIP traffic. We observed poor VoIP performance on a large number of ISP backbone networks under favorable end-system configurations. Action for improving today's VoIP performance to reach toll-quality standards, can be taken inside the network and at the receiver. Inside the network, our findings make a strong case for marking and identifying the voice traffic, in order to give it preferential treatment. At the receiver, it is important that the playout buffer scheme, should be carefully chosen to match the delay pattern.

Acknowledgments

Special acknowledgment goes to Amit Vyas for helping with the processing of the traces and for contributing with many valuable discussions throughout the project.

We are also grateful to RouteScience for providing the traces; in particular to W. Feick for collecting the measurements; also to M. Lloyd, S. Finn, O. Baldonado for their useful feedback. Thanks to A. Clark from Telchemy, R. Cox and M. Perkins from AT&T, D. Vleeschauwer and F. Poppe from Alcatel for their clarifications on the references. Finally, acknowledgment to D. Cheriton for motivating this study, with useful discussions on the need to identify the voice traffic with its own DiffServ codepoint.

References

[1] P. Brandy, "A technique for investigating on/off patterns of speech", *Bell Labs Tech.Journal*, 44(1):1-22, January 1965.

[2] CISCO Systems, "Playout delay enhancements for VoIP", online documentation *http://www.cisco.com/univercd/cc/td/doc/product/software/ios21/121newft/121t/121t5/dt_pod.html.*

[3] R. Cox, M.Perkins, "Results of a subjective listening test for G.711 with frame erasure concealment", *AT&T contribution to T1A1.7/99-016*, May 1999.

[4] A. Clark, "Modeling the effects of burst packet loss and recency on subjective voice quality", *Proc. of IP Telephony Workshop*, March 2001.

[5] A. Clark, R.Liu, "Comparison of TS 101 329-5 Annex E with PAMS and PSQM", Temp.Doc. 061 for *TIPHON#23*, July 2001.

[6] A.Clark, R.Liu, "Comparison of TS101 329-5 Annex E with Emodel", Temp.Doc. 062 for *TIPHON#23*, July 2001.

[7] ETSI TS 101 329-5, Annex E, "QoS measurement methodologies", November 2000.

[8] ETSI TS 101 329-6 "Actual measurements of network and terminal characteristics and performance parameters in TIPHON networks and their influence on voice quality", July 2001.

[9] France Telecom R&D, "Study of the relationship between instantaneous and overall subjective speech quality for time-varying quality speech sequences: influence of the recency effect", *ITU Study Group 12, contribution D.139*, May 2000.

[10] France Telecom R&D, "Continuous assessment of time-varying subjective vocal quality and its relationship with overall subjective quality", *ITU Study Group 12, Contribution COM 12-94-E*, July 1999.

[11] J. Gruber, L. Strawczynski, "Subjective effects of variable delay and speech clipping in dynamically managed voice systems", *IEEE Trans. on Communications*, vol. 33, No.8, Aug.1985.

[12] *ITU-T* Recommendation G.107, "The Emodel, a computational model for use in transmission planning", December 1998.

[13] *ITU-T* Recommendation G.108, "Application of the Emodel: a planning guide", September 1998.

[14] *ITU-T* Recommendation G.113, "Transmission impairments due to speech processing", February 2001.

[15] *ITU-T* Recommendation G.114, "One way transmission time", May 2000.

[16] W.Jiang, H.Schulzrinne, "Analysis of on-off patterns in VoIP and their effect on voice traffic aggregation", *Proc. of ICCCN 2000.*

[17] W.Jiang, H.Schulzrinne, "QoS measurement of real-time multimedia services in the Internet", *Columbia University Tech. Report* CUCS-015-99.

[18] M.Karam, F.Tobagi, "Analysis of the delay and jitter of voice traffic over the Internet", *Proc. of Infocom 2001.*

[19] N. Kitawaki, K. Itoh, "Pure delay effects on speech quality in telecommunications", *IEEE Journal on Selected Areas in Communications*, vol . 9, no.4, May 1991.

[20] T. Kostas, M. Borella, I. Sidhu, G. Schuster, J. Grabiec, "Real-time voice over packet-switched networks", *IEEE Network* January/February 1998.

[21] Y. Liang, N. Farber, B. Girod, "Adaptive playout scheduling using time-scale modification in packet voice communications", *Proc. of ICASSP* 2001.

[22] A. Markopoulou, F. Tobagi, "An adaptive playout that optimizes the perceived VoIP quality", *work in progress.*

[23] S. Moon, J. Kurose, D. Towsley, "Packet audio playout delay adjustment: performance bounds and algorithms", ACM/Springer Multimedia Systems, vol. 6, pp.17-28, January 1998.

[24] C. Perkins, O. Hodson, V. Hardman, "A survey of packet loss recovery techniques for streaming audio", *IEEE Network*, Sept./Oct. 1998.

[25] R. Ramachandran, J. Kurose, D. Towsley, H. Schulzrinne, "Adaptive playout mechanisms for packetized audio applications in wide-area networks", *Proc. of Infocom 1994.*

[26] H. Schulzrinne, online class notes from "Advanced internet systems" on the "public switched telephone system", *http://www.cs.columbia.edu/~hgs/teaching.*

[27] K. Sriram, W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data", *IEEE JSAC*, vol. 4, (6):833-846, Sept.1986.

[28] UCL, Department of Computer Science, Robust Audio Tool (RAT), *http://ww-mice.cs.ucl.ac.uk/multimedia/software/rat/*

[29] V. Vleeschauwer, J. Janssen, G. Petit, F. Poppe, Alcatel Technical Report "Quality bounds for packetized voice transport", *Alcatel Technical Report*, 1st Quarter 2000.

[30] S. Voran, "Speech quality of G.723.1 coding with added temporal discontinuity impairments", *Proc. of ICASSP* May 2001.

[31] J. Walker, J. Hicks, "Evaluating data networks for VoIP", *NetIQ white paper*.